

CLASSIFICATION OF *MALWARE* FAMILIES USING *NAÏVE* *BAYES* CLASSIFIER

Skripsi

Diajukan Untuk Memenuhi
Persyaratan Guna Meraih Gelar Sarjana
Informatika Universitas Muhammadiyah Malang



RAMADAN PRATAMA
(201410370311253)

Data Science

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MUHAMMADIYAH MALANG
2021

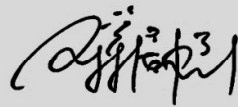
LEMBAR PERSETUJUAN
CLASSIFICATION OF *MALWARE* FAMILIES USING *NAÏVE BAYES* CLASSIFIER

Ramadan Pratama
(201410370311253)

Telah Direkomendasikan Untuk Diajukan Sebagai
Judul Tugas Akhir Di
Program Studi Informatika Universitas Muhammadiyah Malang

Menyetujui,

Dosen I



Denar Regata Akbi S.Kom., M.Kom.
NIP. 108.1612.0591

Dosen II



Vinna Rahmayanti S S.Si., M.Si
NIP. 108.3060.71990

LEMBAR PENGESAHAN

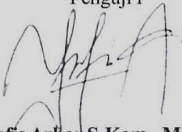
Classification of Malware Families Using Naïve Bayes Classifier

Diajukan untuk memenuhi
Persyaratan Guna Meraih Gelar Sarjana Strata I
Program Studi Informatika Universitas Muhammadiyah Malang

Ramadan Pratama
(201410370311253)

Menyetujui,

Penguji I


Yufis Azhar S.Kom., M.Kom.
NIP. 10814100544

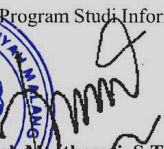
Penguji II



Fauzi Dwi Setiawan Sumadi ST., M.CompSc.
NIP. 180307061992

Mengetahui,




Rita Indah Marthasari, S.T, M.Kom
NIP. 10806110422

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini:

NAMA : Ramadan Pratama
NIM : 20141037031153
FAKULTAS/JURUSAN : TEKNIK / TEKNIK INFORMATIKA

Dengan ini saya menyatakan bahwa Tugas Akhir dengan Judul “**Classification of Malware Families Using Naïve Bayes**” beserta seluruh isinya adalah karya saya sendiri dan bukan merupakan karya tulis orang lain, baik sebagian maupun seluruhnya, kecuali dalam bentuk kutipan yang telah disebutkan sumbernya.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila kemudian ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya ini, atau ada klaim dari pihak lain terhadap keaslian karya saya ini maka saya siap menanggung segala bentuk risiko/sanksi yang berlaku.

Malang, 8 Juli 2021

Yang Membuat Pernyataan



Ramadan Pratama

Mengetahui,

Dosen I

Dosen II

Denar Regata Akbi S.Kom., M.Kom.
NIP. 10816120591

Vinna Rahmayanti S.Si., M.Si
NIP. 108306071990

KATA PENGANTAR

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Dengan memanjatkan puji syukur kehadiran Allah SWT atas limpahan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan tugas akhir yang berjudul “***Classification of Malware Families Using Naïve Bayes***”.

Tulisan ini disajikan pokok-pokok bahasan mengenai klasifikasi *malware family* menggunakan metode *Naïve Bayes* mulai dari rancangan, cara kerja, hingga pengembangannya. Selain itu juga dijelaskan mengenai implementasi serta pengujian dari klasifikasi menggunakan Naïve Bayes

Peneliti menyadari sepenuhnya bahwa dalam penulisan tugas akhir ini masih banyak kekurangan dan keterbatasan. Oleh karena itu, peneliti sangat mengharapkan saran yang membangun agar tulisan ini bermanfaat untuk perkembangan ilmu pengetahuan kedepan.



DAFTAR ISI

LEMBAR PERSETUJUAN	ii
LEMBAR PENGESAHAN	iii
LEMBAR PERNYATAAN.....	iv
ABSTRAK.....	v
<i>ABSTRACT</i>	vi
LEMBAR PERSEMBAHAN	vii
KATA PENGANTAR	ix
DAFTAR ISI.....	x
DAFTAR GAMBAR.....	xii
DAFTAR TABEL.....	xiv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan Penelitian.....	3
1.4 Batasan Masalah.....	3
BAB II TINJAUAN PUSTAKA	4
2.1 <i>Malware</i>	4
2.2 Machine learning.....	4
a. Supervised Learning.....	4
b. Unsupervised Learning.....	4
2.3 Preprocessing	4
2.4 Pearson correlation coefficient.....	5
2.5 Naïve Bayes.....	6
2.6 <i>K-fold cross validation</i>	7
BAB III METODE PENELITIAN	8
3.1 Rancangan Penelitian	8

3.2	Analisa Kebutuhan Sistem	9
3.2.1	Kebutuhan perangkat keras (hardware)	9
3.2.2	Kebutuhan perangkat lunak (software)	10
3.3	Skema Implementasi dan Pengujian Klasifikasi Malware Family	10
3.4	Dataset Malware Family	10
3.5	Preprocessing data dengan Metode Pearson Product-moment Correlation	13
3.6	Split atau Pembagian Data	13
3.7	Klasifikasi menggunakan Naïve Bayes	14
3.8	Evaluasi hasil dari klasifikasi <i>Naïve Bayes</i>	14
BAB IV	IMPLEMENTASI DAN PENGUJIAN	16
4.1	Implementasi	16
4.1.1	Preprocessing Data	16
4.1.2	Implementasi <i>Pearson Correlation Coefficient</i>	16
4.1.3	Klasifikasi dengan metode Naïve Bayes	19
4.2	Pengujian	23
4.2.1	Pengujian klasifikasi <i>Naïve Bayes</i> dengan komposisi data <i>training</i> dan data <i>testing</i> berbeda	24
4.2.2	Perbandingan akurasi <i>Naïve Bayes</i> dengan <i>K fold cross validation</i> menggunakan dataset sebelum dan sesudah <i>feature selection</i>	28
4.2.3	Perbandingan hasil klasifikasi dengan penelitian sebelumnya	30
BAB V	PENUTUP	31
5.1	Kesimpulan	31
5.2	Saran	31
	Daftar Pustaka	32

DAFTAR GAMBAR

Gambar 2.1 positive correlation.....	5
Gambar 2.2 negative correlation.....	5
Gambar 2.3 no correlation	5
Gambar 2.4 perhitungan pearson correlation coefficient.....	5
Gambar 2.5 formula pearson's correlation coefficient	6
Gambar 2.6 formula naive bayes	6
Gambar 2.7 10-fold cross validation.....	7
Gambar 3.1 skema rancangan penelitian	8
Gambar 3.2 skema implementasi dan pengujian	10
Gambar 3.3 contoh dataset CICInvesAndMal2019	11
Gambar 3.4 klasifikasi menggunakan Naïve Bayes	14
Gambar 4.1 proses input library.....	17
Gambar 4.2 proses pemisahan atribut independen	17
Gambar 4.3 proses pembagian dataset CICInvesAndMal2019	17
Gambar 4.4 proses pearson correlation coefficient.....	17
Gambar 4.5 korelasi antar atribut.....	18
Gambar 4.6 proses penghapusan fitur yang tidak digunakan	19
Gambar 4.7 proses import dataset malware	20
Gambar 4.8 proses pemisahan atribut independen	20
Gambar 4.9 isi dari atribut family.....	21
Gambar 4.10 proses pembagian dataset menjadi data training dan data testing.....	21
Gambar 4.11 proses klasifikasi Naïve Bayes.....	22
Gambar 4.12 klasifikasi data testing.....	22
Gambar 4.13 proses menghitung nilai akurasi dengan confusion matrix	23
Gambar 4.14 diagram pembagian data training 60% dan data testing 40%	24
Gambar 4.15 hasil klasifikasi malware 60% data training dan 40% data testing	24
Gambar 4.16 diagram pembagian data training 70% dan data testing 30%	25
Gambar 4.17 hasil klasifikasi malware 70% data training dan 30% data testing	25
Gambar 4.18 diagram pembagian data training 80% dan data testing 20%	26
Gambar 4.19 hasil klasifikasi malware 80% data training dan 20% data testing	26
Gambar 4.20 diagram pembagian data training 90% dan data testing 10%	27
Gambar 4.21 hasil klasifikasi malware 90% data training dan 10% data testing	27
Gambar 4.22 grafik hasil pengujian naïve bayes	28



DAFTAR TABEL

Tabel 3.1 kebutuhan perangkat keras (hardware)	9
Tabel 3.2 kebutuhan perangkat lunak (software).....	10
Tabel 3.3 malware family	11
Tabel 4.1 hasil pengujian naïve bayes	27
Tabel 4.2 perbandingan hasil akurasi dari implementasi k-fold cross validation.....	29
Tabel 4.3 perbandingan RF, k-NN, Naïve Bayes	30



Daftar Pustaka

- [1] D. J. Wu, C. H. Mao, T. E. Wei, H. M. Lee, and K. P. Wu, "DroidMat: Android malware detection through manifest and API calls tracing," *Proc. 2012 7th Asia Jt. Conf. Inf. Secur. AsiaJCIS 2012*, pp. 62–69, 2012, doi: 10.1109/AsiaJCIS.2012.18.
- [2] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-An, and H. Ye, "Significant Permission Identification for Machine-Learning-Based Android Malware Detection," *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3216–3225, 2018, doi: 10.1109/TII.2017.2789219.
- [3] A. H. Lashkari, A. F. A. Kadir, L. Taheri, and A. A. Ghorbani, "Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification," *Proc. - Int. Carnahan Conf. Secur. Technol.*, vol. 2018-October, no. Cic, pp. 1–7, 2018, doi: 10.1109/CCST.2018.8585560.
- [4] Z. Xu, K. Ren, and F. Song, "Android malware family classification and characterization using CFG and DFG," *Proc. - 2019 13th Int. Symp. Theor. Asp. Softw. Eng. TASE 2019*, pp. 49–56, 2019, doi: 10.1109/TASE.2019.00-20.
- [5] M. A. Jerlin and K. Marimuthu, "A New Malware Detection System Using Machine Learning Techniques for API Call Sequences," *J. Appl. Secur. Res.*, vol. 13, no. 1, pp. 45–62, 2018, doi: 10.1080/19361610.2018.1387734.
- [6] L. Liu, B. sheng Wang, B. Yu, and Q. xi Zhong, "Automatic malware classification and new malware detection using machine learning," *Front. Inf. Technol. Electron. Eng.*, vol. 18, no. 9, pp. 1336–1347, 2017, doi: 10.1631/FITEE.1601325.
- [7] L. Massarelli, L. Aniello, C. Ciccotelli, L. Querzoni, D. Ucci, and R. Baldoni, "Android malware family classification based on resource consumption over time," *Proc. 2017 12th Int. Conf. Malicious Unwanted Software, MALWARE 2017*, vol. 2018-Janua, pp. 31–38, 2018, doi: 10.1109/MALWARE.2017.8323954.
- [8] A. R. Yogaswara, D. R. Akbi, V. Rahmayati, and S. Nastiti, "Malware Familiy Classification using k-Nearest Neighbor (k-NN)," vol. 3357, no. 1, pp. 1–5, 2020.
- [9] H. Zhang, C. T. Liu, J. Mao, C. Shen, R. L. Xie, and B. Mu, "Development of novel in silico prediction model for drug-induced ototoxicity by using naïve Bayes classifier approach," *Toxicol. Vitr.*, vol. 65, no. September 2019, 2020, doi: 10.1016/j.tiv.2020.104812.

- [10] P. Chandrasekar and K. Qian, "The Impact of Data Preprocessing on the Performance of a Naïve Bayes Classifier," *Proc. - Int. Comput. Softw. Appl. Conf.*, vol. 2, pp. 618–619, 2016, doi: 10.1109/COMPSAC.2016.205.
- [11] Y. C. Zhang and L. Sakhanenko, "The naive Bayes classifier for functional data," *Stat. Probab. Lett.*, vol. 152, pp. 137–146, 2019, doi: 10.1016/j.spl.2019.04.017.
- [12] M. Sewak, S. K. Sahay, and H. Rathore, "Comparison of Deep Learning and the Classical Machine Learning Algorithm for the Malware Detection," *Proc. - 2018 IEEE/ACIS 19th Int. Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distributed Comput. SNPD 2018*, pp. 293–296, 2018, doi: 10.1109/SNPD.2018.8441123.
- [13] L. Taheri, A. F. A. Kadir, and A. H. Lashkari, "Extensible android malware detection and family classification using network-flows and API-calls," *Proc. - Int. Carnahan Conf. Secur. Technol.*, vol. 2019-Octob, no. Cic, 2019, doi: 10.1109/CCST.2019.8888430.
- [14] T. Pektas, Abdurrahman. Acarman, "Ensemble Machine Learning Approach for Android Malware Classification Using Hybrid Features," *Proc. 10th Int. Conf. Comput. Recognit. Syst. CORES 2017*, 2018, doi: 10.1007/978-3-319-59162-9.
- [15] B. Raju and R. Bonagiri, "A cavernous analytics using advanced machine learning for real world datasets in research implementations," *Mater. Today Proc.*, no. xxxx, pp. 4–7, 2020, doi: 10.1016/j.matpr.2020.11.089.
- [16] H. Fernando and J. Marshall, "What lies beneath: Material classification for autonomous excavators using proprioceptive force sensing and machine learning," *Autom. Constr.*, vol. 119, no. June, p. 103374, 2020, doi: 10.1016/j.autcon.2020.103374.
- [17] E. C. Blessie and E. Karthikeyan, "Sigmis: A feature selection algorithm using correlation based method," *J. Algorithms Comput. Technol.*, vol. 6, no. 3, pp. 385–394, 2012, doi: 10.1260/1748-3018.6.3.385.
- [18] J. D. Chee, "Pearson's Product-Moment Correlation: Sample Analysis," *ResearchGate*, no. May 2015, 2016, doi: 10.13140/RG.2.1.1856.2726.
- [19] Z. Zakeri, N. Mansfield, C. Sunderland, and A. Omurtag, "Cross-validating models of continuous data from simulation and experiment by using linear regression and artificial neural networks," *Informatics Med. Unlocked*, vol. 21, no. July, p. 100457, 2020, doi:

10.1016/j.imu.2020.100457.

- [20] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A new sampling method in particle filter based on Pearson correlation coefficient," *Neurocomputing*, vol. 216, pp. 208–215, 2016, doi: 10.1016/j.neucom.2016.07.036.
- [21] J. D. Chee and T. Queen, "Pearson's Product Moment Correlation: Sample Analysis," *ResearchGate*, no. May 2015, 2016, doi: 10.13140/RG.2.1.1856.2726.
- [22] M. Singh, M. Wasim Bhatt, H. S. Bedi, and U. Mishra, "Performance of bernoulli's naive bayes classifier in the detection of fake news," *Mater. Today Proc.*, no. xxxx, 2020, doi: 10.1016/j.matpr.2020.10.896.
- [23] S. Saud, B. Jamil, Y. Upadhyay, and K. Irshad, "Performance improvement of empirical models for estimation of global solar radiation in India: A k-fold cross-validation approach," *Sustain. Energy Technol. Assessments*, vol. 40, no. June, p. 100768, 2020, doi: 10.1016/j.seta.2020.100768.
- [24] G. Jiang and W. Wang, "Error estimation based on variance analysis of k-fold cross-validation," *Pattern Recognit.*, vol. 69, pp. 94–106, 2017, doi: 10.1016/j.patcog.2017.03.025.



UNIVERSITAS MUHAMMADIYAH MALANG
FAKULTAS TEKNIK
PROGRAM STUDI TEKNIK INFORMATIKA
 Jl. Raya Tlogomas 246 Malang 65144 Telp. 0341 - 464318 Ext. 247, Fax. 0341 - 460782

FORM CEK PLAGIARISME LAPORAN TUGAS AKHIR

Nama : Ramadan Pratama

NIM : 201410370311253

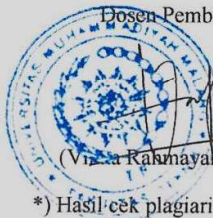
Judul TA : Classification of Malware Families Using Naïve Bayes Classifier

Hasil Cek Plagiarisme dengan Turnitin

No.	Komponen Pengecekan	Nilai Maksimal Plagiarisme (%)	Hasil Cek Plagiarisme (%) *
1.	Bab 1 – Pendahuluan	10 %	5%
2.	Bab 2 – Daftar Pustaka	25 %	3%
3.	Bab 3 – Analisis dan Perancangan	25 %	10%
4.	Bab 4 – Implementasi dan Pengujian	15 %	9%
5.	Bab 5 – Kesimpulan dan Saran	5 %	4%
6.	Makalah Tugas Akhir	20%	2%

Mengetahui,

Dosen Pembimbing



(Vina Rahmayanti S.Si., M.Si)

*) Hasil cek plagiarism bisa diisikkan oleh salah satu pembimbing